Dynamic Online Pricing with Incomplete Information Using Multi-Armed Bandit Experiments

Eric M. Schwartz*

Kanishka Misra†

Jacob Abernethy‡

September 2016

Abstract

Consider the pricing decision for a manager at large online retailer, such as Amazon.com, that sells millions of products. The pricing manager must decide on real-time prices for each of these product. Due to the large number of products, the manager must set retail prices without complete demand information. A manager can run price experiments to learn about demand and maximize long run profits. There are two aspects that make the online retail pricing different from traditional brick and mortar settings. First, due to the number of products the manager must be able to automate pricing. Second, an online retailer can make frequent price changes. Pricing differs from other areas of online marketing where experimentation is common, such as online advertising or website design, as firms do not randomize prices to different customers at the same time.

In this paper we propose a dynamic price experimentation policy where the firm has incomplete demand information. For this general setting, we derive a pricing algorithm that balances earning profit immediately and learning for future profits. The proposed approach marries statistical machine learning and economic theory. In particular we combine multi-armed bandit (MAB) algorithms with partial identification of consumer demand into a unique pricing policy. Our automated policy solves this problem using a scalable distribution-free algorithm. We show that our method converges to the optimal price faster than standard machine learning MAB solutions to the problem. In a series of Monte Carlo simula-

*Ross School of Business, University of Michigan. Contact: ericmsch@umich.edu

†Rady School of Management, University of California, San Diego. Contact: kamisra@ucsd.edu

‡Department of Electrical Engineering and Computer Science, University of Michigan. Contact: jabernet@umich.edu

tions, we show that the proposed approach perform favorably compared to methods in computer science and revenue management.

1 Introduction

1.1 Overview

Consider the pricing decision for a manager at large online retailer. According to a 2015 estimate, Amazon.com sells 356.2 million products, Jet.com sells 24.6 million products and Walmart.com sells 4.3 million products [1]. In these markets, managers have to set real-time retail prices for each of these products. To learn optimal prices, they may use market research or run real-time experiments.

There are two features of this setting that make it different from brick and mortar retail settings. First, given the number of products sold by large online retailers, pricing decisions must be largely automated. Specifically, it is simply not feasible for a manager to run market research, estimate price elasticity, and set optimal prices for each product. Second, online sellers can vary prices nearly continuously and often randomize those changes for learning purposes (White House, 2015). This differs from a traditional retail setting where retailers face high costs, known as menu costs, to change prices limiting the number price changes (Anderson et al., 2015).

In this paper we propose a dynamic price experimentation policy where the firm has incomplete demand information. We maintain the objective of price experimentation is to maximize long-run profits. We treat pricing decisions as a continuous experiment, so the firm is concerned about profits during that adaptive experimentation. By contrast, consider how a firm may typically test prices. The firm may test a set of prices, one at a time, observe demand and profits at each level, and then select the price that led to highest profits. This approach is intuitive and may be seen in practice, but it is also a benchmark in the operations research literature. For example, the algorithm proposed in Besbes and Zeevi (2009), represents the standard non-parametric method. This approach is best described as learn-then-earn.

Instead of two distinct phases of learning and earning, we frame this same problem as a dynamic optimization problem with the goal of maximizing *earning while learning*. Our proposed pricing algorithm sequentially sets prices to balance currently earning profits and learning about demand for future profits. At

[1] `https://learn.scrapehero.com/how-many-products-does-jet-com-sell/` and `https://learn.scrapehero.com/how-many-products-does-walmart-com-sell-vs-amazon-com/` [Accessed on 17 March 2016]

any time, the firm sets a price and observes which consumers were exposed to the prices and their purchase decisions. The firm incorporates this additional information for future pricing decisions.

We resolve this earning-while-learning problem by employing multi-arm bandit (MAB) methods. The MAB problem is a fundamental dynamic optimization problem in reinforcement learning, an area within machine learning (Sutton and Barto, 1998). The problem has a long history in statistics and operations research (Gittins et al., 2011). [2] In the MAB problem, the decision maker is faced with a set of possible decisions, also known as 'arms.' The term 'arm' historically comes from the slot machine's alternative name, the 'one-armed bandit'. Typically, each arm has stable reward distribution unknown to the decision maker, but she has to select arms with the goal of optimizing cumulative reward. In the learning-earning tradeoff (also known as, explore-exploit), the value of uncertainty in a reward's mean is due to potential future gains from learning. But there are many variations of bandit problems as well as a variety of solution approaches.

Our proposed solution, departs from common MAB implementations, as we account for the two key features of the online retail pricing problem setting: wide variety of products and real-time changes for those millions of products. To ensure our model can be used for a wide variety of products, we make minimal assumptions about the underlying demand curve for a particular product. Instead of assuming each product's demand curve to come from the same family of distributions, we opt for an approach as flexible as possible. This is what we mean by *robust* and *non-parametric*. By limiting parametric assumptions, our algorithm will be robust to any arbitrary unknown true demand system.

Our algorithm builds on the Upper Confidence Bound (UCB) algorithm in the non-parametric MAB literature (Auer et al., 2002). The family of UCB algorithms work as follows. In each time period, the algorithm assigns each arm a so-called UCB value – the sum of expected reward and potential value from experimentation. Then the algorithm plays the arm with the highest value. The decision maker observes a noisy reward, and updates these values for each arm. The UCB algorithm is guaranteed to be 'best' non-parametric algorithm for any bounded payoff function in terms of achieving the theoretically optimal rate of error in a finite-time setting (Lai, 1987). Among many MAB algorithms available, UCB is the one best suited for non-parametric setting to be extended for this application.

We extend this MAB algorithm to incorporate economic theory. In the typical UCB algorithm, when a particular price is charged (an arm is played), the firm's observations are limited to profits (reward) from

[2] For a recent overview of MAB in the marketing literature, see Schwartz et al. (2016).

that price (arm). So its learning is limited to expected profits. Formally, we assume that a consumer's choice structure satisfies the weak axiom of revealed preference (WARP). With this additional yet minimal assumption, when a consumer is exposed to any price, the manager can *partially identify* the consumer's underlying preference across different potential prices. For example, if a consumer purchases a good at $\$3$, the manager can infer she would have purchased at any price below $\$3$. And if a consumer does not purchases a good at $\$3$, the manager can infer she would not have purchased at any price above $\$3$.

We consider a setting where each consumer makes a single purchase decision, so we rely on cross-sectional identification to estimate aggregate demand. We assume each consumer belong to a single segment, in practice these segments are based on both demographics and usage patterns. Unlike traditional retail segmentation, due to the abundance of data, there can be a large number of segments for online retailers. For example, Google analytics and Facebook analytics offer over 1,000 segments to advertising brands. [3] Our model of consumer heterogeneity is flexible and semi-parametric. We only estimate a parameter that reflects within-segment variation in valuation (willingness to pay). If consumers within a segment all have different preferences that parameter will be large, and our method will allow large within-segment heterogeneity. Specifically, this parameter only reflects the size of the range of valuations in the segment. The heterogeneity across segments is non-parametric, as the midpoint of partially identified intervals vary across segments. Segments can be based on any number of observed customer characteristics.

Since we make minimal assumptions about the shape or structure of demand model, our algorithm is robust to heterogeneity, context, and selection and can be used for a variety of products. This includes situations where profit functions are discontinuous or multi-modal. It even allows for special price effects (e.g., 9-endings and contextual factors). This could also include situations with observable selection (across segments) and unobserved selection (within a segment). For example, consider a new product, consumers in the early-adopter segment are more likely to visit the website soon after launch, and their valuations would likely be higher than those who adopt later.

In addition to handling a wide variety of products, another key feature of online retail is pricing in real-time at large scale. We ensure our algorithm can run in real-time for millions of products. Our proposed algorithm has minimal estimation and samples prices at speeds orders of magnitude faster than current solutions. In a similar two-period version of our pricing problem, the solution in Handel and Misra (2015)

[3] see segmentation suggestions for app developers on Facebook `https://www.facebook.com/help/analytics/1568220886796899/` and Google analytics users `https://support.google.com/analytics/answer/3123951?hl=en`.

suggests first period prices in about 15 hours of computation time for one product. Building on the algorithms in reinforcement learning (Auer, 2002), our proposed method can calculate about 2 million prices per minute. Therefore, it could be used for real-time pricing by online retail websites.

1.2 Contributions

With the emergence of big data, we see an increase in machine learning applications in marketing (Chintagunta et al., 2016). But a natural critique is machine learning algorithms' absence of economic theory. This work illustrates how we can bridge this gap. We propose a novel combination of economic theory with machine learning. To marketing and economics, we bring these scalable reinforcement learning methods to expand the types of addressable dynamic optimization problems. While a two-period solution was available, the method could not realistically scale with time. We propose a fast and scalable algorithm rooted in theory.

To machine learning, we introduce distribution-free theory of demand to improve existing algorithms theoretically and empirically. Typically the models of active learning in computer science often rely on stylized demand models since they are amenable to formal analysis. But they may lack the economic theory that, as we find, can improve the optimal pricing algorithms. Beyond a wide range of Monte Carlos simulations across 12 pricing algorithms and 5 demand curve scenarios, we also provide theoretical support. We prove our proposed algorithm's performance using a novel analysis technique for MAB algorithms. This uses a potential function, which departs from the standard methods for formal analyses of finite-time minimax regret.

1.3 Related Literature

1.3.1 Literature on Pricing

Our paper adds to a large literature on pricing and learning demand in marketing, economics, and operations research. We highlight the important differences between our paper and the current literature. Much of the current literature makes strong assumptions about the information the firm has about demand of each product. Most of the literature in marketing and operations assumes that firms make product pricing decisions based on knowing the demand curve (see for example Oren et al. (1982); Rao and Bass (1985); Wernerfelt (1986); Smith (1986); Bayus (1992); Rajan et al. (1992); Acquisti and Varian (2005); Nair (2007);

Akcay et al. (2010); Jiang et al. (2011)). These methods assume that the firm has access to perfect information about the demand curve and consider the optimal dynamic pricing given this information. We argue that it is infeasible for large online retailers to know the demand curve for millions of products.

A second related literature assumes that firms know demand only up to a parameter (Rothschild, 1974; Lodish, 1980; Aghion et al., 1991; Braden and Oren, 1994; Kalyanam, 1996; Biyalogorsky and Gerstner, 2004; Bergemann and Valimaki, 1996; Aviv and Pazcal, 2002; Hitsch, 2006; Desai et al., 2010; Bonatti, 2011; Biyalogorsky and Koenigsberg, 2014). The modeling approach in these papers assumes that the manager knows the structure of demand and learns the parameters. This could be a two-period model (Biyalogorsky and Koenigsberg, 2014) or an infinite-time model (Aghion et al., 1991). In the infinite-time model, these papers assume Bayesian updating. Aghion et al. (1991) consider a very general model where the manager knows the structure of demand up to a parameter (θ), the firms sets prices and observes market outcomes. In subsequent periods the firm updates the posterior belief distributions for the parameters, and then the firm sets prices. They show that under this structure learning can be "inadequate" in cases where the profit function is multi-modal or discontinuous. Inadequate learning is defined as where the agent never acquires adequate knowledge (i.e., asymptotically, adequate with probability zero). Adequate knowledge is defined when the agent knows enough about the true profit function to achieve ex-post optimal profits. Aghion et al. (1991) concludes: "even when learning goes on forever, it does not result in adequate knowledge" (pg. 642).

We argue that it is important for the robust pricing policy to incorporate all possible demand curves. Therefore, we make weaker assumptions than assuming a demand model and derive optimal prices. In the econometric sense, by making only weak assumptions about the demand curve, we sacrifice precision for credibility (Manski, 2005). This is consistent with the saying that it is not feasible for a firm to have precise and credible demand information about every single product, so we have to make a trade-off between precision and credibility.

Our non-parametric method builds on the robust pricing literature (Bergemann and Schlag, 2008, 2011; Handel et al., 2013; Handel and Misra, 2015). The dynamic pricing in this literature provides a solution only for a two-periods version of our problem (Handel and Misra, 2015). We differ from these papers by building a solution based on the MAB literature in computer science. The advantage of our model is that it allows for continuous learning and real-time changes to suggested prices. In our paper, we solve the problem for any number of periods using scalable distribution-free algorithms.

The current work also contributes to some theoretical work in operations research, namely dynamic pricing and revenue management. The area of revenue management has a large literature that considers dynamic pricing (see Elmaghraby and Keskinocak (2003); den Boer (2015) for a reviews of the literature). Much of this work assumes a functional form for the demand curve and uses Bayes updating on its parameters. Our work falls into a category known as dynamic pricing without inventory constraints, where dynamics are due to incomplete information and learning. Within this literature our paper fits with the less studied and more recent stream of non-parametric work. Non-parametric approaches, exemplified by Besbes and Zeevi (2009), consider pricing policies in an incomplete information setting. Here the authors consider algorithms that minimize the maximum ex-post statistical regret from not charging the optimal static price. The proposed algorithm divides the the sales horizon into an "exploration" phase during which the demand function is learned and an "exploitation" phase during which the estimated optimal price is used. The firm has to ex-ante set the length of experimentation stage. More recent additions to this literature Wang et al. (2014) and Lei et al. (2016) improve the convergence results. The algorithms proposed in these paper also consider distinct phases for exploration then exploitation, or as we refer to it, "learning then earning.". Instead, in our paper we the learning and earning phases simultaneously, accounting for the the potential value from learning. This is consistent with the broader stream of the MAB literature.

1.3.2 Literature on Multi-Armed Bandits

We consider the problem of pricing using MAB methods, which are not typically used for pricing, but do stretch across computer science, statistics, operations research, and marketing. The MAB problem is the quintessential problem of the field of reinforcement learning literature in computer science (for an overview, see Sutton and Barto (1998)).

A large part of this literature provides theoretical analysis and mathematical guarantees of algorithms. The algorithms are policies to adaptively select the arms to achieve the best profits. The objective function for these algorithms is to minimize the statistical regret. *Statistical regret* "is the loss due to the fact that the globally optimal policy is not followed all the times" Auer et al. (2002). That is the difference between the achieved profits and the ex-post optimal profits, if the decision maker knew the true average profits for each arm. Algorithms are compared based on the bounds on regret. This bound represents the worst case performance, the maximum possible regret for any possible distribution of the true rewards across the arms.

These UCB policies provide the backbone of a stream of MAB solutions in reinforcement learning

coming from statistical learning (Agrawal, 1955; Auer et al., 2002). Lai and Robbins (1985) first obtained these "nearly optimal index rules in the finite-horizon case" where the indices can be interpreted as "upper confidence bounds for the expected rewards," hence UCB (Brezzi and Lai (2002), pp. 88-89). While these index rules do not provide the exactly optimal solution to the optimization problem with discounted infinite sum of expected rewards solved by the Gittins index, these rules are asympotitcally optimal for arbitrarily large finite-time horizons, T. As $T \to \infty$, the UCB-based index rule achieve optimal performance with respect to maximizing the expected sum of rewards through T periods "from both the Bayesian and frequentist viewpoints for moderate and small values of [T]" (Brezzi and Lai (2002), pp. 88-89).

Asymptotic theory links the finite-horizon undiscounted case and the infinite-horizon discounted multi-armed bandit problem (Brezzi and Lai, 2002). Depending on the discount factor, $disc$, the UCB directly approximates the Gittins index (Lai, 1987). When setting $T = (1 - disc)^{-1}$, as $disc \to 1$, then the UCB in Lai (1987) (pg. 1113) is not only asympotitcally optimal for the finite-horizon undiscounted problem but also for the infinite-horizon undiscounted problem from Gittins. The link is strengthened as Brezzi and Lai (2002) derive an Approximate Gittins Index (AGI), with a structure, the sum of expected reward and an exploration bonus, exactly the same as that of the UCB.

We add to this literature by allowing non-parametric (partially identified) learning across the potential prices. Formally, we assume that individual indirect utility functions are non increasing in prices, this must be true for any demand function that satisfies the weak axiom of revealed preference (WARP) (see proposition 3.D.3 in Mas-Colell et al. (1995), and discussed in the next section).

We note if the decision maker is willing to make stronger parametric assumptions about the demand curve, alternative streams of MAB algorithms based on parametric models are more appropriate. One such algorithm is the earliest Bayesian formulation of the MAB problem, which lead to Thompson Sampling Thompson (1933). A more prominent formulation with Bayesian learning lead to the Gittins index (Gittins, 1989) and Whittle index (Whittle, 1980).

Later developments focus on the so-called parametric bandit setting. These approaches combine bandit algorithms with regression models.[4] In such algorithms the expected rewards across arms can be correlated based on some known similarity of arms. For instance, a the UCB combines generalized linear model to form UCB-GLM algorithm (Dani et al., 2008; Filippi et al., 2010; Rusmevichientong and Tsitsiklis, 2010).

[4]This type of parametric MAB problem goes by many names, linear bandit or attribute-based bandit, and is related to other MAB variants, including contextual bandit or bandit with side information.

Thompson Sampling also applies to any (generalized) linear model or even hierarchical model Schwartz et al. (2016); Scott (2010). Even a Gittins-like index has been extended for correlations among arms. Price is just one attribute describing naturally correlated arms.

The learning-and-earning problem for pricing relates to a broader class of problems, optimizing marketing experiments or so-called A/B testing. The emerging framework is using multi-armed bandits to optimally balance earning while learning. These approaches most commonly appear in online advertising or website design (Hauser et al., 2009; Urban et al., 2013; Schwartz et al., 2016).

We argue that pricing is different from other marketing decisions, in two key ways. First, economic theory gives strong predictions that individual indirect utility functions are non increasing in prices, this is not true for other marketing decisions. Without particular parametric assumptions, learning about one ad creative or website design does not inform predictions of others. Second, unlike advertising, the randomization of prices is imperfect. Retailers do not commonly offer the same product to different consumers at a given point in time. [5] Therefore, price changes can happen only across time and not across consumers, this could lead to potential selection concerns.

2 Dynamic Multi-Period Monopoly Pricing

2.1 Model setup and maintained assumptions

In our setup we assume the firm is a monopolist and sets its online prices to maximize profits for a constant marginal cost product. In this section, we state our main assumptions in our analysis. We first discuss our assumptions on the demand side and then our assumptions on the supply side.

We assume there are a large set of potential consumers with unit demand for each product. For each consumer we assume the following:

1. she has stable preferences,

2. she has a stable budget over time,

3. she faces a stable outside option, and

[5] Amazon.com has run price experiments in 2000 and due to consumer feedback release a statement say "random testing was a mistake, and we regret it because it created uncertainty and complexity for our customers, and our job is to simplify shopping for customers. That is why, more than two weeks ago, in response to customer feedback, we changed our policy to protect customers" `http://cnnfn.cnn.com/2000/09/28/technology/amazon/`.

4. her choice structure satisfied the weak axiom of reveal preference (WARP).

With these assumptions, we represent the consumer's preference as v_i. In any purchase occasion, when facing a price p, her indirect utility can be written as $u_i = v_i - p$ and will purchase the good if and only if $u_i \geq 0$, that is, $v_i \geq p$. The assumption of stable preference guarantees that v_i does not change over time, this rules out learning (Erdem and Keane, 1996), stockpiling (Hendel and Nevo, 2006), network externalities (Nair et al., 2004), reference price effect (Kalyanaram and Winer, 1995; Winer, 1986) and strategic consumers (Nair, 2007). Unlike much of the prior work on dynamic prices (e.g., Nair (2007), in our paper firms change prices for learning the demand curve as opposed to inter-temporal price discrimination.

Our next set of assumptions consider the heterogeneity across consumers. We assume each consumer i can be assigned to an ex-ante segment s. In practice, these segments maybe based on observable variables, both demographics and behavioral patterns, or model-based criteria. Unlike traditional retail segmentation, online retailers have the ability to use a large number of segments, for example Google analytics allows different 1,000 segments. [6] We assume the firm knows the aggregate proportion of consumers in each segment, ψ_s (or uses this based on some previous model-based segmentation).

Let v_i be the preference for consumer i, and let v_s represent the midpoint of range of consumer valuations within segment s. We let within-segment heterogeneity of valuations to be δ. That is, the preference of all consumers in a segment are within δ of the segment midpoint, v_s. Taken together,

$$v_i \in [v_s - \delta, v_s + \delta] \forall i \in s.$$

We emphasize the generality of these assumptions. Within each segment, we allow for any distribution of preferences within this range; therefore, we note that v_s is not assumed to be the mean or the median of the segment valuations. Across segments, We allow for fully non-parametric heterogeneity across segments. This assumption allows cross-sectional learning of consumer preferences. We note that we will estimate δ in our empirical algorithm. This can be viewed as a measure of "quality of segmentation". If the firm's ex-ante segmentation does not group consumers with similar preferences, then the estimate of δ will be large. But if the firm's ex-ante segmentation does group consumers with similar preference, then δ will be small.

On the supply side, we assume that the firm is a monopolist who sets online prices to maximize profits

[6] We note in practice such segmentation is used by any advertisers using data from Facebook `https://www.facebook.com/help/analytics/1568220886796899/` or Google analytics `https://support.google.com/analytics/answer/3123951?hl=en`, Accessed March 2016.

for a constant marginal cost product. The main deviation we make from the standard pricing literature (e.g., Oren et al. (1982); Rao and Bass (1985)) we assume that firm does not know consumer valuations. We assume that the only information available to the firm at the time of initial pricing is that consumer valuations are between $[v_L, v_H]$. The interpretation here is that if the product is sold for v_L (can be zero) all consumers will purchase for sure, and if the product is sold for v_H, no consumers will buy. Consistent with the robust pricing literature (Bergemann and Schlag, 2008, 2011; Besbes and Zeevi, 2009; Handel et al., 2013; Wang et al., 2014; Handel and Misra, 2015; Lei et al., 2016) we assume within this range the firm does not know the distribution of consumer preferences across or within segments. Our motivation for this assumption is that it is infeasible for the manager to have credible priors for millions of products.

We assume that the firm does not price discriminate across consumers. Formally, the firm observes a consumer's identity and segment membership, but only after the consumer's makes a purchase decision. If we had full information, there exists an optimal static price that a monopolist would charge. However, due to the lack of information the monopolist must experiment with prices. We assume that the firm can change prices quickly. White House (2015) reports that Amazon.com can change prices within 15 minutes. In our model, we will assume that prices can change after every N consumers who visit the product. [7]

2.2 Overview of multi-armed bandit

We begin by formulating the pricing problem as a dynamic optimization problem. We assume there exist a finite set of K prices that the firm can chose from $p \in \{p_1, \ldots, p_K\}$. For any price, p, the firm faces an unknown true demand $D(p)$. We assume a constant marginal cost (set to zero for ease of exposition). Given this setup the true profit is given by $\pi(p) = pD(p)$.

In our empirical analysis, we assume the firm does not know the true profit function $\pi(p)$, instead the firm observes realizations of profits for each price p_k. Suppose by time t, the firm has charged p_k a total of n_{kt} times. Let $\pi_{k,1}, \pi_{k,2}, \ldots \pi_{k,n_{kt}}$ be realizations from each time price p_k has been charged. We assume that these are drawn from an unknown probability distribution with a mean at the true profit $\pi(p_k)$. We refer to the sample mean at time t as $\bar{\pi}_{kt} = \frac{\sum_{\tau=1}^{n_{kt}} \pi_{k\tau}}{n_{kt}}$. By definition, we must have $\sum_{k=1}^{K} n_{kt} = t$.

A pricing *policy or algorithm*, ϕ, selects prices based on the history of past prices and earned profits.

[7] Websites `http://camelcamelcamel.com` and `https://thetracktor.com` we can track price changes. For example, we tracked the following prices Apr 13, 2016 06:20 PM $1.59; Apr 13, 2016 12:44 AM $4.75; Apr 12, 2016 03:56 PM $5.10; Apr 12, 2016 07:02 AM $1.59; Apr 11, 2016 10:15 PM $5.12 for a Prismacolor Scholar Colored Pencil Sharpener on Amazon.com

Mathematically this can be described as, $p_t = \phi(\{p_\tau, \pi_\tau | \tau = 1, \ldots, t-1\})$. The policy maps data from all previous experiments onto price.

To evaluate a policy's performance, we follow the literature and use *regret*. The key criterion to evaluate policies is minimizing maximum regret (i.e., minimax regret). Regret for a policy is defined as the expected profit loss due to not always playing the unknown ex-post optimal profit-maximizing fixed price (Lai and Robbins, 1985). The notion of regret is standard in the computer science and statistical machine learning literature on MAB problems (Lai and Robbins, 1985; Auer, 2002; Schwartz et al., 2016). In the statistical decision theory literature the decision criterion is called "minimax regret" ((Berger, 1985) pp 376). This was first proposed by Wald (1950) and has been axiomatized in the economic literature (Milnor, 1954; Stoye, 2011). This criterion has been used to study pricing in economics (Bergemann and Schlag, 2008, 2011; Handel et al., 2013) and marketing (Handel and Misra, 2015). The economic interpretation of regret is the "forgone profits" due to price experimentation.

We note regret is appropriate because of the active learning setting. We need an "ex-ante" criteria to evaluate a pricing policy. By "ex-ante" we mean, the objective function must be one that can be calculated without knowing the true demand curve. This is necessary as we will evaluate different pricing policies before the polices are implemented and true profits are realized. Specifically we cannot consider "ex-post" criteria such as total profits, as this cannot be used to evaluate a policy *before* true profits are realized. In our paper, we propose an algorithm and then mathematically establish the upper bound on the possible regret. This is an "ex-ante" evaluation, for any possible data-generating process the regret from our algorithm must be below this upper bound.

Formally, we represent regret as the distance to the optimal profits. We define the ex-post profit maximizing price to be p^* for all t yielding an expected profit $\mu^* = \mathbb{E}[\pi(p)] = p^* D(p^*)$ each time period. The regret of a policy ϕ through time t is

$$\text{Regret}(\phi, t) = \mathbb{E}\left[\sum_{\tau=1}^{t} \pi^* - \pi_\tau\right] = \sum_{\tau=1}^{t} (\pi^* - \pi_{p_\tau}) = \pi^* t - \sum_{k=1}^{K} \pi(p_k) \mathbb{E}[n_{kt}]$$

where π_t is profit realized in time period t.

When considering analysis of regret, we does not observe true realizations of profits ($\pi_1, \ldots, \pi_K$, and therefore π^*). The analysis instead considers all possible realizations of $\pi_1, \ldots, \pi_K$ as this is know before running the algorithm. Next we consider the possible realization that generates the "worst case" or maximum

regret for give policy ϕ. The economic interpretation of this is to consider a feasible demand curve ($D(p)$) that results in the maximum regret given a pricing policy ϕ. The best algorithm is one that can minimize the maximum regret.

In the MAB literature, the minimax regret optimal solution for this non-parametric problem is a policy involving an index rule scoring each action with its UCB of expected rewards (Agrawal, 1955; Auer, 2002; Lai and Robbins, 1985; Lai, 1987). This policy is proven to be the asymptotically best possible performance in terms of achieving the lowest maximum regret.

The structure of the index assigned to each action in the UCB algorithm is the sum of expected reward and an exploration bonus. For instance, in the focal algorithm in Auer (2002), UCB1, the index for action k at time T is based on only the sample mean reward of the arm and an exportation bonus. In our notation this translates to

$$\text{UCB1}_{kt} = \bar{\pi}_{kt} + \sqrt{\frac{\alpha \log t}{n_{kt}}}$$

then the arm with the highest UCB value is selected to be played in the next round. The parameter $\alpha > 0$ can be tuned to the particular scenario, but a choice of $\alpha = 2$ suffices to obtain a quite general bound on regret. We will discuss the details of UCB in greater detail in Section 3.

The amount by which UCB exceeds the expected reward is called the exploration bonus, representing the value of information. The particular structure inside the exploration bonus, $2\log(t)/n_{kt}$, follows from the structure of proof of the algorithm's optimality. To prove that it is asymptotically optimal, the value of information (exploration bonus) is defined to ensure that cumulative regret (the difference between the cumulative reward and the optimal cumulative reward) grows slowly at logarithmic rate in time, with arbitrarily high probability. We illustrate this in our proof for our algorithm later in the theoretical analysis.

Auer (2002) notes that the UCB1 algorithm only considers the number of times each action is played and does not account for the variance in outcomes from the trial of each arm. They provide an additional algorithm called UCB-Tuned which they report better performance, though they are no able to analytically derive the regret bounds. For this algorithm they define

$$\mathrm{V}_{kt} = \left(\frac{1}{n_{kt}} \sum_{\tau=1}^{n_{kt}} \pi_{k\tau}^2 \right) - \bar{\pi}_{kt}^2 + \sqrt{\frac{2 \log t}{n_{kt}}}$$

$$\text{UCB-TUNED}_{kt} = \bar{\pi}_{kt} + \sqrt{\frac{\log t}{n_{kt}} \min\left(\frac{1}{4}, \mathrm{V}_{kt}\right)}$$

An assumption in non-parametric multi-armed bandit algorithms, including the UCB algorithms, is that the profit outcomes in any two actions are uncorrelated. That is, the realized profits when action k is played does not inform us of the possible profits with another action j is played. While in many marketing applications, this is a valid assumption depending on the design of the experiment, such as, website design (Hauser et al., 2009). And in other marketing applications with correlated actions, parametric assumptions are required to capture those correlations, as shown in online advertising (Schwartz et al., 2016).

Pricing is different. In an application to pricing, however, we can add non-parametric demand learning (Handel et al., 2013) to MAB algorithms. We will prove the regret convergence rates with adding demand learning for the UCB1 algorithm and will then adapt the UCB-Tuned algorithm to account for variance in observed profits. In the next section we discuss how we can add non-parametric demand learning and then will discuss an updated model.

We note that while we account for demand learning, our model is different to dynamic minimax regret problem discussed in Handel and Misra (2015). In our model we account for the fact that observed outcomes of a prior price experiment can impact expected demand for all other price points. However we do not consider endogenous learning when considering the exploration bonus for current price experiments. We note we consider a very different context to Handel and Misra (2015) who consider a context where all consumers to be exposed to every price experiment. Instead we consider online prices that can change rapidly and only a few consumers are exposed to prices. The benefit of not including endogenous learning is analytical tractability. In a two-period pricing problem the solution in Handel and Misra (2015) suggests first period prices in about 15 hours of computation time for one product. Our proposed method take about one minute to calculate 2 million prices. Therefore can be used in real-time by online retail websites.

2.3 Learning demand curve from price experiments

In this section we will discuss how we the researcher can learn preferences across different price experiments. In this section our key parameter of interest is the demand for each product at each price level, or $D(p_k)\ \forall k \in [1, K]$. This section is based on the demand side assumptions we make in section 2.1. The implication of these assumptions is that if a consumer is willing to purchase a product a price p_1, she will be willing to purchase the product for any price $p_2 < p_1$. Similarly, if the consumer does not purchase the product at p_2, she will not be willing to purchase the product for any price $p_1 > p_2$. [8]

Formally we can define a set of possible consumer preference as $\Theta \equiv \{\theta_1, ..., \theta_K\}$, where θ_k refers to a preference that satisfies: (a) $\theta_k - p > 0$ for all $p \leq p_k$ and (b) $\theta_k - p < 0$ for all $p > p_k$. Alternatively, θ_k represents a preference under which the highest amount the consumer will purchase this good for is p_k. With our assumption of WARP these reduce to (a) $\theta_k - p_k > 0$ and (b) $\theta_k - p_{k+1} < 0$. The set Θ is discrete and finite as the set of possible prices is discrete and finite.

If we consider products that are purchased repeatedly, this we can use this information to identify bounds for each consumers valuations. Consider an example of where we observe the following price experiments for a consumer i. She purchases at \$3, does not purchase at \$8, purchases at \$2 and does not purchase at \$6. We can say that the true preference for this consumer (v_i) must be between \$3 and \$6.

Formally, define $d_{ik} = 1$ if the consumer purchases at price p_k and $d_{ik} = 0$ if the consumer does not purchase. We can define the identified set of types for consumer i as $H[\theta_i] = \{\theta_l : \max_k(d_{ik} = 1) \leq l \leq \min_k(d_{ik} = 0)\}$. We can then aggregate this non-parametrically across all consumers to identify the set to all feasible demand curves (see Handel et al. (2013) Section 2 for details).

Requiring many repeat-purchase data for each customer may be overly restrictive or not suited for most online products, so we focus cross sectional learning with segmentation.

2.3.1 Estimating segment level demand

The firm has data on $n_{s,t}$ price experiments for segment s through time t. In each experiment a consumer i in segment s is exposed to a price p_k and makes a purchase decision. If the consumer purchases her valuations are consistent with her valuation. We now extend this section to learning about segment level demand. In this section we will first describe how one can estimate v_s and δ from the observed price

[8] In the treatment choice literature (Manski, 2005) this corresponds to monotone treatment response

experiment, and then we will address how we can estimate bounds for demand, $D(p_k)\ \forall k \in [1, K]$.

For any price p_k we can define the set of valuations (defined as H[$D(p_k)_{s,t}$]) that is consistent with that price as follows: (a) $D(p_k)_{s,t} = 0$ is consistent with consumers being of types $\{\theta_1, ..., \theta_{p_k-1}\}$, or types where consumers will not purchase at price p_k; (b) $D(p_k)_{s,t} = 1$ is consistent with consumers being of types $\{\theta_k, ..., \theta_{p_K}\}$, or types where consumers will purchase for sure at price p_k; (c) $D(p_k)_{s,t} \in (0, 1)$ is consistent with a mixture of consumer types $\{\theta_1 ... \theta_{p_k-1}\}$ and $\{\theta_k, ..., \theta_{p_K}\}$, or types where some consumers will purchase and other consumers will not purchase.

For any segment we can define p_s^{min} as the highest price where all consumer in segment s purchase. Mathematically, $p_{s,t}^{min} \equiv \max\{p_k | D(p_k)_{s,t} = 1\}$. And define p_s^{max} as the lowest price where no consumer from segment s purchased. Mathematically, $p_{s,t}^{max} \equiv \min\{p_k | D(p_k)_{s,t} = 0\}$.

Consider the following example to describe our estimation. Suppose we have data for segment 1 with the following observations:

- At a price $\$1$, 100% of consumers purchased
- At a price $\$2$, 100% of consumers purchased
- At a price $\$3$, 50% of consumers purchased
- At a price $\$4$, 0% of consumers purchased
- At a price $\$5$, 0% of consumers purchased

Given these data we would define $p_1^{min} = \$2$ and $p_1^{max} = \$4$ as given our data for this segment we know for sure all consumers purchase at prices $\$2$ or lower and no consumers will purchase as a price at a price higher than $\$4$.

We know that given the information so far $\forall i \in s, v_{i,s} \in [p_{s,t}^{min}, p_{s,t}^{max}]$. We can define an estimated mid-point of the segment valuations (v_s) and the segment level $\delta_{s,t}$ as

$$\hat{v}_{s,t} = \frac{p_{s,t}^{max} + p_{s,t}^{min}}{2}$$
$$\hat{\delta}_{s,t} = \frac{p_{s,t}^{max} - p_{s,t}^{min}}{2}$$

The interpretation of $\delta_{s,t}$ here is that it is the smallest δ that can rationalize the observed decisions for consumers in segment s after t observed price experiments. We note that this estimate will be consistent,

that is in the limit as $t \to \infty$ (and there is enough price variation, we identify the true δ (call this δ^* for each segment. Formally, we have $\lim_{t\to\infty} P(\hat{\delta}_{st} = \delta^*) = 1$. However these will be biased for any finite t, $\delta_{s,t}$ will be biased downwards. In order to correct for this bias we use the assumption that $\delta_s = \delta$, that is all segments have the same δ. Our methodology to estimate the small sample bias follows Handel et al. (2013).

In any time period t, consider the set $\{\hat{\delta}_{1t}, ..., \hat{\delta}_{St}\}$. We then estimate the maximum of that set,

$$\hat{\delta}_t = \max\{\hat{\delta}_{st}, s \in S\}.$$

This will be also be biased downwards relative to δ^*. Again, $\hat{\delta}_t$ would be consistent for δ^*, we follow the econometric literature on estimating boundaries to correct for this bias in our estimator (see Karunamuni and Alberts (2005) for a review). Denote the bias as γ. Our estimator is similar to that used in Handel et al. (2013), which is an adaption of the Hall and Park (2002) estimator. [9] Define $f(.)$ as the empirical distribution of $\hat{\delta}_{st}$ (controlling for segment $s$ize) across S for fixed t. Formally $f(x) = \sum_{s\in S} \psi_s \mathbf{1}(\hat{\delta}_{st} = x)$. Note incorporating ψ_s in our estimate for f allows us to account for the fact that different segments are of different sizes.

Our estimator for γ is:

$$\hat{\gamma}_t = \sum_{\hat{\delta}_{st} \in \Delta_t} (\hat{\delta}_t - \delta_{st}) f(\hat{\delta}_{st})$$

This estimator is consistent as $\lim_{t\to\infty} \hat{\gamma}_t = 0$, by our assumption of a common δ^* across segments. Handel et al. (2013) show that $\hat{\delta}_t + \hat{\gamma}_t$ provides a reliable and conservative estimate for the true δ^*.

Define $\hat{v}^{\min}_{s,t} = \hat{v}_{s,t} - (\hat{\delta}_t + \hat{\gamma}_t)$ and $\hat{v}^{\max}_{s,t} = \hat{v}_{s,t} + (\hat{\delta}_t + \hat{\gamma}_t)$ to represent the lowest and highest possible consumer valuations within segment s. The key output from this analysis is for each segment of consumers s, we can identify the identified set of consumer preference $H_t[v_{i,s}]$ as follow:

$$H_t[v_{i,s}] = [\hat{v}^{\min}_{s,t}, \hat{v}^{\max}_{s,t}] = [\hat{v}_{s,t} - (\hat{\delta}_t + \hat{\gamma}_t), \hat{v}_{s,t} + (\hat{\delta}_t + \hat{\gamma}_t)]$$

[9] Formally Hall and Park (2002) boundary estimator considers a setup where the econometrician observes N draws from a continuous univariate distribution F with a unknown and finite upper boundary. The Handel et al. (2013) estimator is a discrete analog to these methods, since the distribution of $\hat{\delta}_{st}$ is discrete.

2.4 Learning demand curve at population level

Using distribution-free partial identification, aggregated to the population-level, we gain information to narrow the set of possible demand curves. As we accumulate data of demand for different prices, we aim to bound expected demand (and expected profit) more tightly. After gaining new data, we can update the bounds. For each price p_k, the true demand is $D(p_k)$. Without any data we can define the identification region $H[(p_k)]$ as $H[D(p_k)] = [0, 1]$. Here we will use the identified set of valuations within each segment to estimate the bounds on aggregate demand and profits.

The aggregate demand at a price p_k is the number of consumers in each segment that have valuations $v_{i,s} \leq p_k$. Define $F_s(.)$ to be the true cumulative density of all valuation with a segment s. Then, we can rewrite aggregate demand as,

$$D(p_k) = \sum_{s \in S} \psi_s (1 - F_s(p_k))$$

where ψ_s is the (known) proportion of consumers in segment s.

However, we do not observe $F_s(p_k)$. Therefore we can consider bounds of this distribution. From our estimation in the previous section, we know that $F_s(p_k) = 0$ if p_k is less than the lower bound of valuations for segment s, $\hat{v}_{s,t}^{\min}$. Similarly $F_s(p_k) = 1$ if p_k is greater than the upper bound of valuations for segment s, $\hat{v}_{s,t}^{\max}$. Therefore, we can define the identified region for demand at price p_k as

$$H[D(p_k)|t] = [\sum_{s \in S} \psi_s \mathbf{1}\left(\hat{v}_{s,t}^{\min} \leq p_k\right), \sum_{s \in S} \psi_s \mathbf{1}\left(\hat{v}_{s,t}^{\max} \leq p_k\right)].$$

This aggregation is best described in an example. Suppose we have two segments of equal sizes. For segment 1 say have identified preferences to be between $[\$2, \$6]$ and for segment 2 we have identified preference to be between $[\$5, \$9]$. We can identify the feasible demand sets for prices $\$1$ to $\$10$ as follows:

- $H[D(\$1)] = H[D(\$2)] = [1, 1]$ (point identified), as consumers in segment 1 and segment 2 will purchase for sure
- $H[D(\$3)] = H[D(\$4)] = H[D(\$5)] = [0.5, 1]$, as consumers in segment 1 may or may not purchase and consumers in segment 2 will purchase for sure
- $H[D(\$6)] = [0, 1]$, as consumers in segment 1 and segment 2 may or may not purchase purchase

- $H[D(\$7)] = H[D(\$8)] = H[D(\$9)] = [0, 0.5]$, as consumers in segment 1 will not purchase and consumers in segment 2 may or may not purchase
- $H[D(\$10)] = [0, 0]$, as consumers in segment 1 and segment 2 will not purchase

Using the identified profit, we define profit bounds for each price as, $H[\pi(p_k)|t] = p_k H[D(p_k)|t]$. This gives us the lower and upper bound of true profit after t observations, which we define as $LB(\pi(p_k), t)$ and $UB(\pi(p_k), t)$. In summary, we have

$$\begin{aligned} H[\pi(p_k)|t] =& [LB(\pi(p_k), t), UB(\pi(p_k), t)] \\ LB(\pi(p_k), t) =& p_k \sum_{s \in S} \psi_s \mathbf{1}\left(\hat{v}_{s,t}^{\min} \leq p_k\right) \\ UB(\pi(p_k), t) =& p_k \sum_{s \in S} \psi_s \mathbf{1}\left(\hat{v}_{s,t}^{\max} \leq p_k\right) \end{aligned}$$

3 Upper confidence bound with learning partially identified demand (UCB-LPI)

In this section, we extend the UCB1 algorithm to accommodate profit maximization by incorporating learning demand with partial identification. We define this upper confidence bound bandit algorithm incorporating learning partially identified demand (UCB-PI).

The UCB-PI (untuned) index is,

$$\textbf{UCB-PI}\text{-untuned}_{kt} = \begin{cases} \bar{\pi}_{kt} + p_k \sqrt{\frac{2 \log t}{n_{kt}}} & \text{if } UB(\pi(p_k), t) > \max_l(LB(\pi(p_l), t)) \\ 0 & \text{if } UB(\pi(p_k), t) \leq \max_l(LB(\pi(p_l), t)) \end{cases} \tag{1}$$

There are two key differences between our proposed algorithm and the UCB1 algorithm described Auer (2002). First, we assign an action a non-zero value only if the upper bound of potential returns are higher than the highest lower bound across all action. In an partial identification sense, we only consider a an action if it is not dominated by another action. From an economic sense, there is no reason to explore an action, if we know an alternative action will lead to higher profits with certainty. Empirically, we will examine how the set of active prices varies over time, eliminating dominated prices and focusing on a set including the true optimal price.

Second, we scale the exploration bonus by price p_k. This is because we know $D(p_k) \in [0, 1]$, and therefore $\pi(p_k) \in [0, p_k]$. But the original UCB1 algorithm was defined where each action's reward had

the same potential range, e.g., $[0, 1]$, regardless of action. By restricting demand, we impose a natural upper bound of profit that depends on price.

In following section, we first prove properties of the UCB-PI and show that regret is lower than UCB1. Then, we define a tuned version of the UCB algorithm analogous to the UCB-tuned algorithm in Audibert et al. (2009).

3.1 Theoretical performance

In this section we provide theoretical guarantees for the **UCB-PI** index. To derive a UCB-style algorithm and prove that it is asymptotically optimal, the value of information (exploration bonus) is reverse engineered to ensure that cumulative regret grows slowly, i.e. at rate logarithmic in the time T, that holds with arbitrarily high probability. This log-regret bound was first shown by Lai and Robbins (1985) for a particular stylized multi-armed bandit problem. Our proof is based on an alternative view of the original UCB analysis. Our techniques involve the use of of potential function described below. This allows us to directly compare our proposed algorithm with **UCB-1**.

The proof we present here differs from the standard UCB proof from Auer et al. (2002) because we use an argument based on *potential function*, which we define in the proof. This argument in the proof is a novel application of these tools for formally analysis of algorithms. We use this alternative approach, in part, because it permits a more general description of the exploration bonus.

For this proof, we use more general notation beyond price and profit to describe actions and rewards. Price p_k played at time t is the action described by $\mathcal{A}^t$. Let π_k at price k is described a random variable R_i^t of reward for $i \in 1, ..., K$.

Let Q_i be a distribution on the reward R_i^t, with support on $[0, p_i]$. Then let the rewards $R_i^1, \ldots, R_i^T \overset{\text{iid}}{\sim} Q_i$, where mean $\mathbb{E}[R_i^t] = \mu_i$. We assume that the largest μ_i is unique and, without loss of generality, assume that the coordinates are permuted in order that μ_1 is the largest ex-post mean reward. Define $\Delta_i := \mu_1 - \mu_i$ for $i = 2, \ldots, K$.

The *bandit algorithm* is a procedure that chooses an action $\mathcal{A}^t$ on round t as a function of the set of past observed action/reward pairs, $(\mathcal{A}^1, R^1_{\mathcal{A}^1}), \ldots, (\mathcal{A}^{t-1}, R^{t-1}_{\mathcal{A}^{t-1}})$.

On round t, the past data are summarized by the count, $N_i^t := \sum_{\tau=1}^{t-1} \mathbb{I}[\mathcal{A}^\tau = i]$, and the empirical mean estimator, $\hat{\mu}_i^t := \frac{\sum_{\tau=1}^{t-1} \mathbb{I}[\mathcal{A}^\tau = i] R^\tau_{\mathcal{A}^\tau}}{N_i^t}$.

Much of the literature and techniques used to analyze finite time multi-armed bandit problems rely on

a standard set of tools known as *deviation bounds* or *concentration inequalities*. Deviation bounds are used to reason about tail probabilities of averages of iid random variables and martingales, for instance. Perhaps the most basic deviation bound is Chebyshev's Inequality, which says that for any random variable X with mean μ and variance σ^2 we have $\Pr\left(|X-\mu| > k\sigma\right) \leq \frac{1}{k^2}$. More advanced results are based on the *Chernoff bounds*, which provide much sharper guarantees on the decay of the tail probability. For example, the *Hoeffding-Azuma Inequality* (Cesa-Bianchi and Lugosi, 2006), which we present below, gives a probability bound on the order of $\exp(-k^2)$, which is much faster than $1/k^2$.

Let us assume we are given a particular deviation bound that provides the following guarantee,

$$\Pr\left(|\mu_i - \hat{\mu}_i^t| > p_i\epsilon \;\middle|\; N_i^t \geq N\right) \leq f(N,\epsilon), \tag{2}$$

where $f(\cdot,\cdot)$ is a function, continuous in $\epsilon > 0$ and monotonically decreasing in both parameters, that controls the probability of a large deviation. While UCB relies specifically on the Hoeffding-Azuma inequality, for now we leave the deviation bound in a generic form.

We define a pair of functions that allow us to convert between values of ϵ and N in order to guarantee that $f(N,\epsilon) \leq \nu$ for a given $\nu > 0$. To this end define

$$\begin{aligned}
\Lambda(\epsilon,\nu) &:= \min\{N \geq 1 : f(N,\epsilon/2) \leq \nu\}, \\
\rho(N,\nu) &:= \begin{cases} \inf\{\epsilon : f(N,\epsilon) \leq \nu\} & \text{if } N > 0; \\ 1 & \text{otherwise,} \end{cases}
\end{aligned}$$

We omit the ν in the argument to $\Lambda(\cdot), \rho(\cdot)$. Note the property that $\rho(N,\nu) \leq \epsilon/2$ for any $N \geq \Lambda(\epsilon,\nu)$.

Note that $\hat{\delta}$, which plays a role in the lower and upper bounds on reward, does not enter this proof, yet we can conclude the proof applies to our proposed algorithm. Indeed, δ is not known to the researcher and must be estimated. Consider the worst case, where segmentation is useless, then $\delta \to 1$. Then the credible intervals for every segment's feasible profit still are the entire possible range. This is the case presented in the proof here. But in practice, $0 \leq \delta \leq 1$, and δ can be smaller than its maximum value, making segmentation useful, and narrowing the partially identified intervals. Therefore, the proposed **UCB-PI** algorithm does no worse than the performance described here.

3.1.1 Bounds for the UCB-PI algorithm

Recall that the **UCB-PI** index is defined in Equation 1 by taking the mean estimated reward plus an exploration bonus for each price p_i. The precise form of the exploration bonus derives from the deviation bound, particularly from the form of $\rho(\cdot)$. In other words, for a fixed choice of $\nu > 0$, we can redefine the algorithm as follows:

$$\textbf{UCB-PI Algorithm:} \qquad \text{on round } t \text{ play } \mathcal{A}^t = \arg\max_i \left\{ \hat{\mu}_i^t + p_i \rho(N_i^t, \nu) \right\} \tag{3}$$

A central piece of the analysis relies on the following potential function, which depends on the current number of plays of each arm $i = 2, \ldots, K$.

$$\Phi(N_2^t, \ldots, N_K^t) := 2 \sum_{i=2}^{K} \sum_{N=0}^{N_i^t - 1} p_i \rho(N, \nu)$$

With our notation, the expected regret can be expressed as

$$\mathbb{E}\left[\text{Regret}_T(\text{UCB})\right] = \sum_{t=1}^{\tau} \mu_1 - \mu_{\mathcal{A}^t}$$

Lemma 1. *The expected regret of UCB is bounded as*

$$\mathbb{E}\left[\text{Regret}_T(\text{UCB})\right] \leq \mathbb{E}\left[\Phi(N_2^{T+1}, \ldots, N_K^{T+1})\right] + O(T\nu)$$

Proof. The additional (statistical) regret suffered on round t of UCB is exactly $\mu_1 - \mu_{\mathcal{A}^t}$. From our deviation bound (Equation 2), we know [10]

$$\mathbb{P}\left(\mu_1 \leq \hat{\mu}_1^t + p_i \rho(N_1^t, \nu)\right) \leq 1 - \nu \quad \text{and} \quad \mathbb{P}\left(\hat{\mu}_{\mathcal{A}^t}^t \leq \mu_{\mathcal{A}^t} + p_{\mathcal{A}^t} \rho(N_{\mathcal{A}^t}^t, \nu)\right) \leq 1 - \nu.$$

To analyze the two inequalities above, we let ξ^t be the indicator variable that one of the above two inequalities fails to hold. We chose $\rho(\cdot)$ so that $\mathbb{P}\left[\xi^t = 1\right] \leq 2\nu$.

[10] Here we can see that if δ were less than its maximum value, the partially identified intervals shrink, and the above probabilities are smaller.

Since the algorithms choose arm $\mathcal{A}^t$, we have

$$\hat{\mu}_1^t + p_1\rho(N_1^t, \nu) \leq \hat{\mu}_{\mathcal{A}^t}^t + p_{\mathcal{A}^t}\rho(N_{\mathcal{A}^t}^t, \nu)$$

If we combine the above two equations, and consider the event that $\xi^t = 0$, then we obtain

$$\mu_1 \leq \hat{\mu}_1^t + p_1\rho(N_1^t, \nu) \leq \hat{\mu}_{\mathcal{A}^t}^t + p_{\mathcal{A}^t}\rho(N_{\mathcal{A}^t}^t, \nu) \leq \mu_{\mathcal{A}^t} + 2p_{\mathcal{A}^t}\rho(N_{\mathcal{A}^t}^t, \nu).$$

Even in the event that $\xi^t = 1$ we have that $\mu_1 - \mu_{\mathcal{A}^t} \leq 1$. Hence, $\mu_1 - \mu_{\mathcal{A}^t} \leq 2p_{\mathcal{A}^t}\rho(N_{\mathcal{A}^t}^t, \nu) + \xi^t$.

Finally, we observe that the potential function was chosen so that $\Phi(N_2^{t+1}, \ldots, N_K^{t+1}) - \Phi(N_2^t, \ldots, N_K^t) = 2p_{\mathcal{A}^t}\rho(N_{\mathcal{A}^t}^t, \nu)$. Recalling that $\Phi(0, \ldots, 0) = 0$,

$$\mathbb{E}\left[\text{Regret}_T(\text{UCB})\right] \leq \mathbb{E}\left[\Phi(N_2^{T+1}, \ldots, N_K^{T+1}) + \sum_{t=1}^{T} \xi^t\right] = \mathbb{E}\left[\Phi(N_2^{T+1}, \ldots, N_K^{T+1})\right] + 2T\nu.$$

□

The final piece we need to establish is that the number of pulls N_i^t of arm i, for $i = 2, \ldots, K$, is unlikely to exceed $\Lambda(\Delta_i, \nu)$.

Lemma 2. *For any $T > 0$ we have $\mathbb{E}\left[\Phi(N_2^{T+1}, \ldots, N_K^{T+1})\right] \leq \Phi(\Lambda(\frac{\Delta_2}{p_2}, \nu), \ldots, \Lambda(\frac{\Delta_K}{p_k}, \nu)) + O(T^2\nu)$.*

Proof of Lemma 2. To obtain the inequality of the lemma, define for every $t = 1, \ldots, T$ and $i = 2, \ldots$ the indicator variable ζ_i^t which returns 1 when $\mathcal{A}^t = i$ given that $N_i^t \geq \Lambda(\frac{\Delta_i}{p_i}, \nu)$, and returns 0 otherwise. We can show that $\zeta_i^t = 1$ with probability smaller than 2ν.

Note that if $\mathcal{A}^t = i$ then the upper confidence estimate for i was larger than that of action 1. More precisely, it must be that $\hat{\mu}_i^t + p_i\rho(N_i^t) \geq \hat{\mu}_1^t + p_1\rho(N_1^t)$. For this to occur, either we had (a) a large underestimate on μ_1, that is $\hat{\mu}_1^t + p_1\rho(N_1^t) \leq \mu_1$. Or, (b) we had a large overestimate on μ_i, that is, $\hat{\mu}_i^t + p_i\rho(N_i^t) \geq \mu_1$. It is clear that (a) occurs with probability less than ν by construction of ρ.

To analyze (b), note that $\mu_1 = \mu_i + \Delta_i$, and we are also given that $N_i^t \geq \Lambda(\frac{\Delta_i}{p_i}, \nu)$ which implies that $p_i\rho(N_i^t) \leq \Delta_i/2$.

$$\hat{\mu}_i^t + p_i\rho(N_i^t) \geq \mu_1 \implies \hat{\mu}_i^t \geq \mu_i + p_i\rho(N_i^t)$$

which happens with probability no more than ν. Therefore,

$$\begin{aligned}\mathbb{E}\left[\Phi(N_2^{T+1},\ldots,N_K^{T+1})\right] \leq & \Phi(\Lambda(\frac{\Delta_2}{p_2},\nu),\ldots,\Lambda(\frac{\Delta_K}{p_k},\nu)) + \mathbb{E}\left[\sum_{i=2}^{K}\sum_{t=1}^{T}\zeta_i^t\right] \\ \leq & \Phi(\Lambda(\frac{\Delta_2}{p_2},\nu),\ldots,\Lambda(\frac{\Delta_K}{p_k},\nu)) + 2T^2\nu\end{aligned}$$

□

We are now able to combine the above results for the final bound.

Theorem 3. *If we set $\nu = T^{-2}/2$, the expected regret of UCB is bounded as*

$$\mathbb{E}\left[\text{Regret}_T(\text{UCB})\right] \leq 8\sum_{i=2}^{K}\frac{p_i\log(T)}{\Delta_i} + O(1).$$

Proof. A standard deviation bound that holds for *all* distributions supported on $[0, p_i]$ is the Hoeffding-Azuma inequality (Cesa-Bianchi and Lugosi, 2006), where the bound is given by $f(N,\epsilon) = 2\exp(-2N\epsilon^2)$. Utilizing Hoeffding-Azuma we have $\Lambda(\epsilon,\nu) = \left\lceil \frac{2\log 2/\nu}{\epsilon^2} \right\rceil$ and $\rho(N,\nu) = \sqrt{\frac{\log(2/\nu)}{2N}}$ for $N > 0$. If we utilize the fact that $\sum_{y=1}^{Y}\frac{1}{\sqrt{y}} \leq 2\sqrt{Y}$, then we see that

$$\begin{aligned}\Phi(\Lambda(\frac{\Delta_2}{p_2},\nu),\ldots,\Lambda(\frac{\Delta_K}{p_k},\nu)) &= 2\sum_{i=2}^{K}\sum_{N=1}^{\Lambda(\frac{\Delta_i}{p_i},\nu)}\rho(N,\nu) \\ &= 2\sum_{i=2}^{K}\sum_{N=1}^{\Lambda(\frac{\Delta_i}{p_i},\nu)}\sqrt{\frac{\log(2/\nu)}{2N}} \\ &\leq 2\sum_{i=2}^{K}2\sqrt{\frac{\log(2/\nu)\Lambda(\frac{\Delta_i}{p_i},\nu)}{2}} \\ &= 4\sum_{i=2}^{K}\frac{p_i\log(2/\nu)}{\Delta_i}\end{aligned}.$$

Combining the Lemma 1 and Lemma 2, setting $\nu = T^{-2}/2$, we conclude the theorem. □

Setting $\rho(N,\nu) = \sqrt{\frac{\log(2/\nu)}{2N}}$ into the UCB algorithm in Equation 3 we get the exploration bonus in our proposed algorithm (see equation 1). The bound for regret for this algorithm is strictly lower than Auer (2002) as all p_is are scaled to be lower than 1. Further adding the additional partial identification implies that an arm is played weakly less than $\Lambda(\epsilon,\nu)$ derived by the Hoeffding-Azuma inequality. Consider the proofs

for Lemmas 1 and 2, as arms are "turned off", we get lower deviation bounds. As we discussed above, the number of arms turned off in an empirical application depends on the value of δ. The bounds derived are for δ at its maximum value, where no arms are turned off, and segmentation is not useful. However the empirical performance of our algorithm should improve for any lower values of δ. The theoretical argument holds true as a worst case analysis.

3.2 The UCB-PI-tuned algorithm

We will present a version of our algorithm where we tune the exploration bonus by considering both the variance out observed outcomes and the size of the bound. This is analogous to the UCB1-tuned algorithm presented in Auer et al. (2002). The V_{kt} represents an upper bound on the reward variance (as opposed to mean). It is also equal to its empirical variance plus an exploration bonus,

$$\mathrm{V}_{kt} = \left(\frac{1}{n_{kt}} \sum_{\tau=1}^{n_{kt}} \pi_{k\tau}^2 \right) - \bar{\pi}_{kt}^2 + \sqrt{\frac{2 \log t}{n_{kt}}}.$$

The upper bound on variance enters the UCB of the mean to control the size of the exploration bonus. We add an additional tuning factor, $2\hat{\delta}$, since it is the size of the range for our partially identified intervals. When δ is large, there is more uncertainty; when it is small, the intervals shrink and so does the exploration bonus.

$$\textbf{UCB-PI}\text{-tuned}_{kt} = \begin{cases} \bar{\pi}_{kt} + 2p_k\hat{\delta}\sqrt{\frac{\log t}{n_{kt}} \min\left(\frac{1}{4}, \mathrm{V}_{kt}\right)} & \text{if } UB(\pi(p_k), t) > \max_l(LB(\pi(p_l), t)) \\ 0 & \text{if } UB(\pi(p_k), t) \leq \max_l(LB(\pi(p_l), t)) \end{cases}$$

The final novel aspect of the proposed algorithm is "shutting off" prices that are dominated. Dominated prices have an upper bound that is still worse than at least some other price's lower lower bound, $UB(\pi(p_k), t) \leq \max_l(LB(\pi(p_l), t))$.

4 Empirical performance: Simulation study

We test our proposed algorithm in a series of simulation experiments. These show its robust performance across unknown true distributions of consumer valuations. Each simulation has the same structure of the data-generating process. Customers arrive, observe the price, and purchase if and only if their valuation is greater than the price. After the customers decide, the firm observes which customers arrived (in particular

the consumer's segment) and their choices. Then the firm sets the price for the next period.

In our simulation we consider a firm with $K = 100$ potential prices from \$0 to \$1 in 0.01 increments. The firm can change prices after every period. Each period, 10 consumers visit, and each consumer belongs to one of $S = 1,000$ segments. We draw the segment probabilities (true ψ_s) from a complex on the uniform distribution. The midpoint of each segment's valuation (true v_s) differs across segments following a distribution, which we vary by simulation condition. Importantly, this distribution is the data generating process and is unknown to the researcher, so it is not assumed in the estimation method. Within each segment, consumers' valuations can be $10c$ above or below the segment midpoint (true δ = 0.1).

4.1 The value of partial identification: Comparing of untuned algorithms

First we consider a simulation to show the advantage of adding partial identification to UCB algorithms. Here we consider the profit implications of the UCB-1 untuned algorithm to the UCB-PI-untuned. We note that the computer science literature has noted that tuned algorithms outperform untuned algortims (Auer, 2002), however feel this is an important comparison to show the advantage of adding partial identification. We also use this first simulation to illustrate how the remaining simulations will work.

For this particular simulation, we assume that the true distribution of segment midpoints is a right-skewed beta distribution, beta$(2, 9)$. This will be one of several tested distributions in large Monte Carlo simulations to follow.

We simulate the true valuations for each segment, simulate consumer decisions, and then see how quickly the two algorithms learn the true valuations over time.

The results of our simulation are shown in Figure 1. In Panel A we plot the prices charged (left column) in each of 200,000 time periods. The top chart show the prices in the UCB1 untuned and the bottom chart shows the prices under UCB-PI untuned. We can see the set of prices tested each period narrows, showing that partial identification allows us to reduce the number of prices experimented. The algorithm focuses on prices near the true ex-post optimal price. Due to this faster learning thanks to partial identified demand curve, the UCB-PI algorithm results in higher ex-post profits (right column). Overall the the UCB-PI attains 91% of ex-post optimal profits, while the UCB1 attains 61% of ex-post optimal profits.

To show why the UCB-PI results in higher profits, in Panel B we plot the estimated δ (left column), which represents the heterogeneity of preferences within a segment. In our simulation the true value is $10c$, we show that we do recover this true valuation and consistent with Handel et al. (2013) we find that in early

simulation we estimate a value of δ that is biased upward. This implies that our learning is not biased, however is slower than if we knew the true δ. We plot the percentage of arms that are active (right column), the other "turned off" due to partial identification. In our simulation we estimate that about 45% of arms are active, this allows the algorithm to focus the exploration of demand.

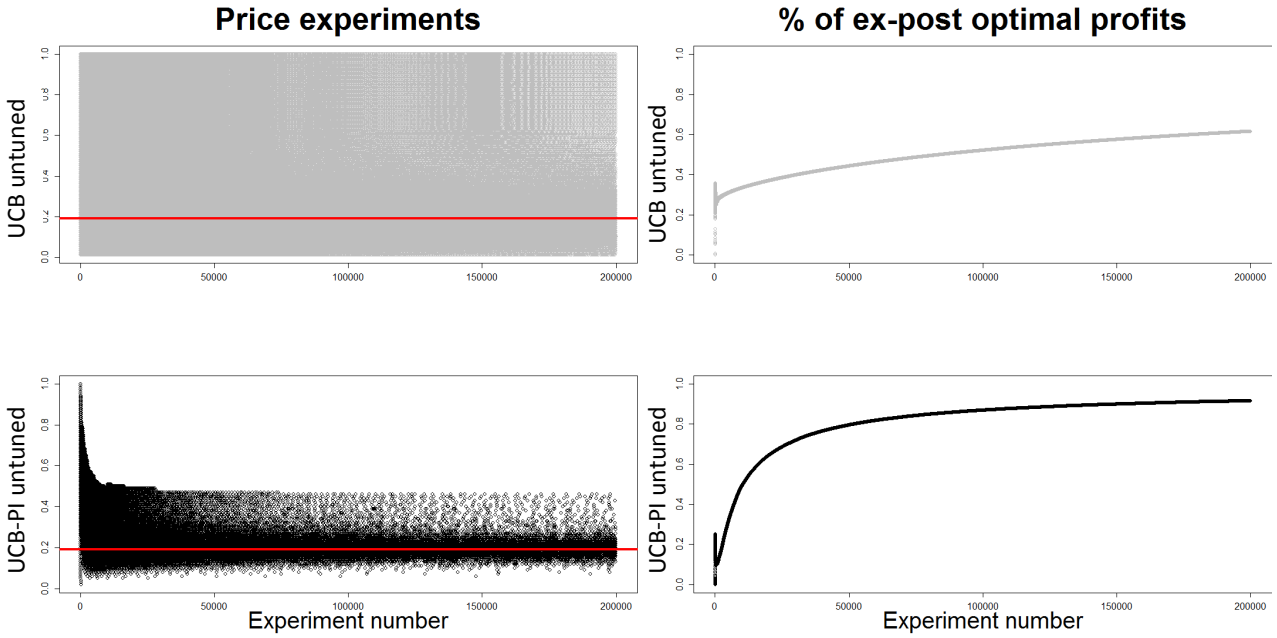
4.2 Comparison of tuned algortims

To compare the tuned algorithms we consider five (5) possible distribution of valuations. These refer to the heterogeneity distribution for v_s. The first three (3) are continuous distributions: a right skewed beta distribution given by beta(2,9), a symmetric beta distribution given by beta(2,2) and a left skewed beta distribution given by beta(9,2). The purpose of these is to consider a range of different possible distributions of consumer preferences. In addition we will consider two (2) bimodal distributions: a bimodal continuous given by beta(0.2,0.3) and a discontinuous finite mixture model with each v_s equal to either \$4 (with 70% chance) or \$9 (30%). From Aghion et al. (1991) we know that Bayesian leaning leads to insufficient learning situations with bimodal continuous and discontinuous profit functions. For our tuned algorithms we will consider 20,000 experimental time periods, with 10 consumers per period.

Figure 2 panel A plots the prices played in each of these five (5) true distributions. Each row of this figure contains the results for a simulation. The first column plots the true ex-post profit function (data-generating process) unknown to the researcher. The second column plot the prices from a UCB-tuned algorithm and the third column plots the prices from our suggested UCB-PI-tuned algorithm. Comparing the price charts with figure 4.1, we do find consistent with the prior literature we do find that tuning leads to less experimentation and a higher focus on profitability. Again, we find that adding partial identification to the UCB algorithm leads to faster learning of demand. Across the simulations, we find that between 50% and 81% of the arms are active. In all settings adding partial identification results in the algorithm setting prices at the optimal levels more often, with more focused experimentation. In the two cases with multimodal demand and profit functions (bimodal continuous and finite mixture), the UCB-PI algorithm does recognize the multiple modes and experiments to find the model with the highest profit.

Turning our attention to profit achieved, in Figure 2 Panel B, we plot the ex-post profit achieved by each algorithm as compared to ex-post optimal profit. The ex-post optimal profit is calculated using the true demand curve and the the firm set the optimal price in every period. We find that across all setting the UCB-PI tuned achieves higher profit that the UCB tuned algorithm. The UCB-PI algorithm achieves above

Panel A: The prices experimented (left column) under UCB-PI (bottom) and UCB (top), both untuned, show how the algorithms learn the true ex-post optimal price (red line). Profit realized (right) plots the cumulative profits realized versus the ex-post optimal.



Panel B: Estimated parameters in the UCB-PI algorthim. The left column consider the estimated max delta, this represents the variation in preferences within a segment. The red line represents the true value used in data genearation. The right column shows the number of arms (prices) considered by the algorithm; the others are turned off due the partially identified profit curves.

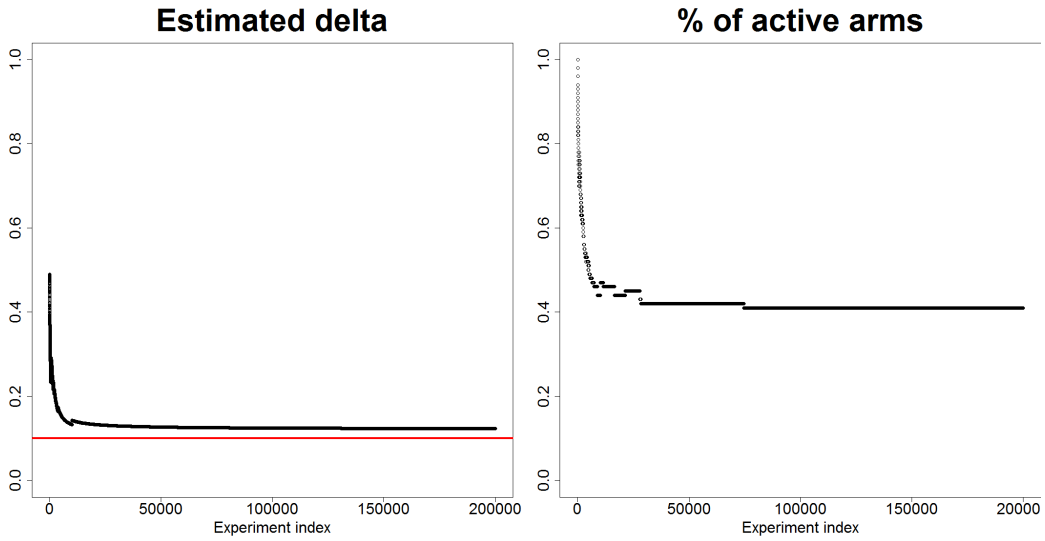


Figure 1: Comparision of the UCB-1 and UCB-PI untuned algorithm for the simulation with true segment valuations from right-skewed distribution, beta(2,9). Overall, adding partial identification increases ex-post profits from 61% to 91%.

95% of ex-post optimal profit in four (4) of the five (5) settings. The finite mixture scenario where learning is difficult (Aghion et al., 1991), the algorithm achieves 89% of ex-post profits.

4.3 Monte Carlo comparisons to alternative algorithms

To consider a broader set of algorithms, we consider the Lean and then Earn algorithm which is consistent with the operations literature (Besbes and Zeevi, 2009). In these algorithms we researcher has to ex-ante set how long the algorithm should learn and how long the algortims should earn. In our experimentation we consider 5 versions of the Learn and then Earn algorithm where learning is set for 0.1%, 1%, 5%, 10% and 25% of experiments. In the computer science literature, Kuleshov and Precup (2014) discuss the advantages of heuristic based models such as e-Greedy. In this model, the algorithm plays the arm with the highest mean profits with a probability 1-ϵ, and a random arm with probability ϵ. Given ϵ is a researcher based input, we consider 3 different values 0.1%, 1% and 5%.

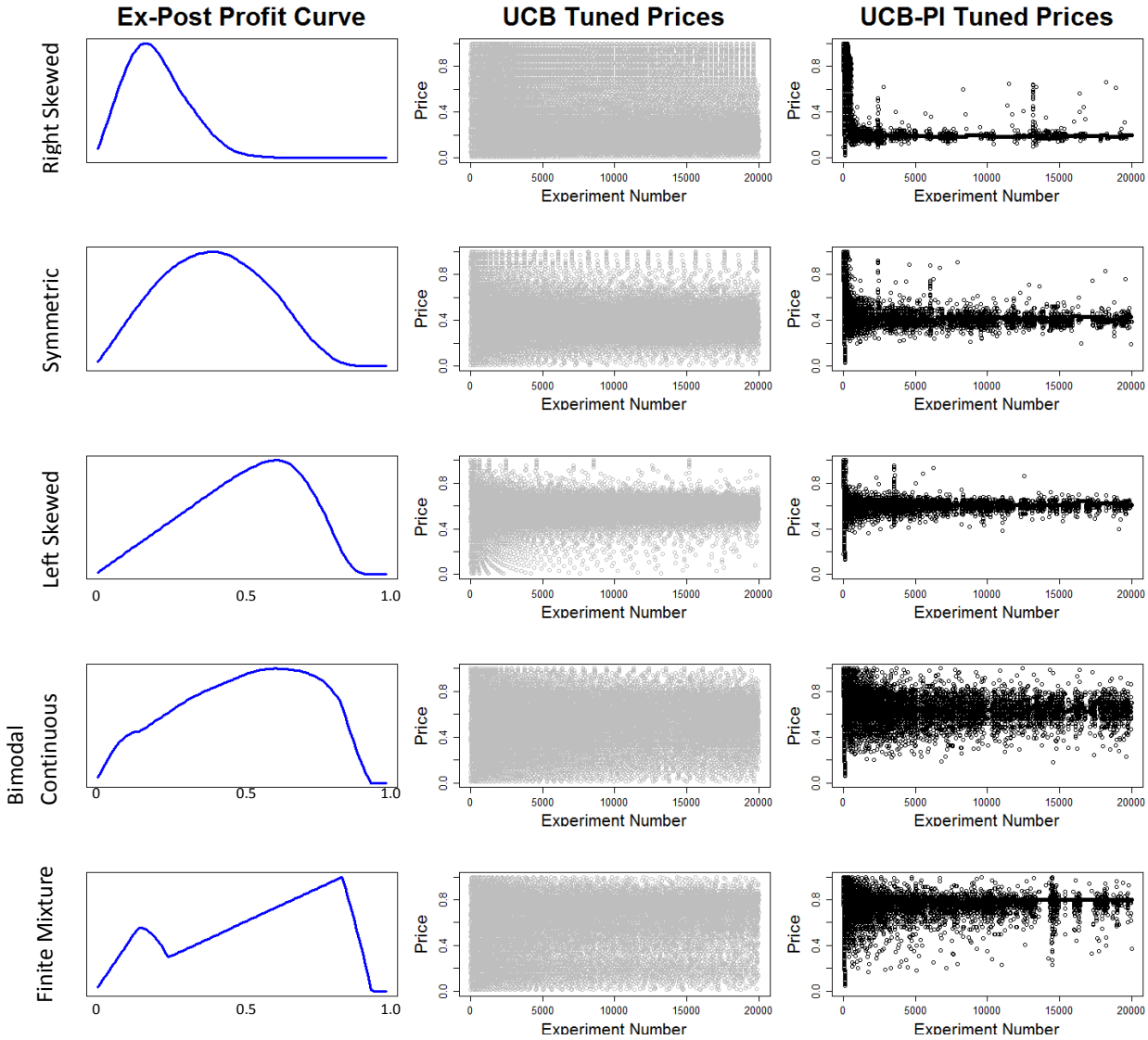
In all we consider 12 different algorithm: UCB untuned and tuned, UCB-PI untuned and tuned, e-greedy with 3 different settings and learn and then earn with 5 different settings. We run 250 different Monte Carlo simulations for each of our 5 different simulation settings. In each simulation run each algorithm for 100,000 time periods with 10 consumers in time period.

The summary of ex-post achieved profits is shown in Figure 3. In panel A (over 2 pages), we plot the ex-post optimal profit for each algorithm. In the first simulation with right skewed preferences, a UCB untuned algorithm achieves 52% (with a variation of 51% and 55% across MC simulations) of ex-post profit. The UCB-PI tuned algorithm achieves on average 98% of ex-post optimal profits. e-Greedy algorithms also achieve 98% of ex-post optimal profits, however with a larger variation across MC simulations. Learn and Earn algorithms achieve between 80% and 94% of ex-post optimal profits depending on the time to learn.

Across all five (5) simulation settings, we find that heuristic based algortims (e-Greedy and lean then earn) do achieve higher ex-post profits than UCB tuned. This finding is consistent with Kuleshov and Precup (2014). However we find that when we add economic theory to the UCB algorithm, the UCB-PI tuned algorithm out-performs or at least is comparable to heuristic based algortims. A key advantage of the theory based algorithms (UCB and UCB-PI) is that they have a lower range of outcome across MC simulations, this is because they are less susceptible to extreme observations.

The ex-post profits across all simulations are summarized in panel B. Here we plot the range (lowest and highest) achieved profit across all We note that while the UCB-PI may not always achieve the highest levels

Panel A: Prices played. The first column contains the true ex-post optimal profit function, the second column contains the UCB Tuned prices and the third column contains the UCB-PI Tuned prices



Panel B: Ex-Post Profits achieved across the five (5) simulation settings

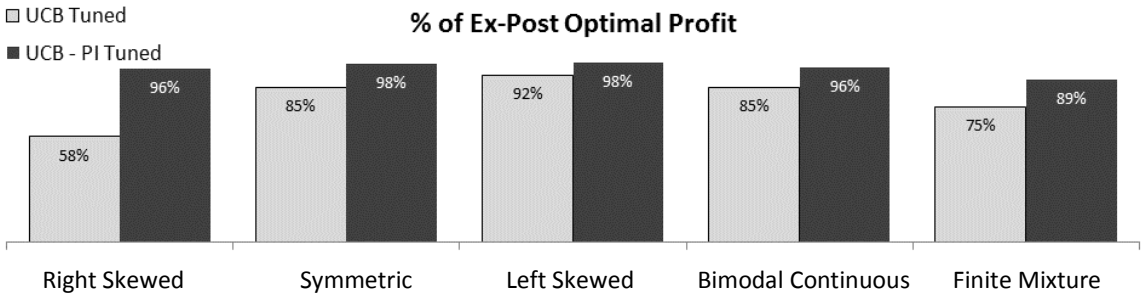


Figure 2: Experimental results under UCB-Tuned and UCB-PI Tuned across five (5) simulation settings.

of profit, however across all simulations the lowest profit achieved by UCB-PI is 92%. This is higher than any other algorithm. This is consistent with the theoretical result the UCB-PI of minimizing the maximum ex-post regret.

5 Conclusion and Future Research

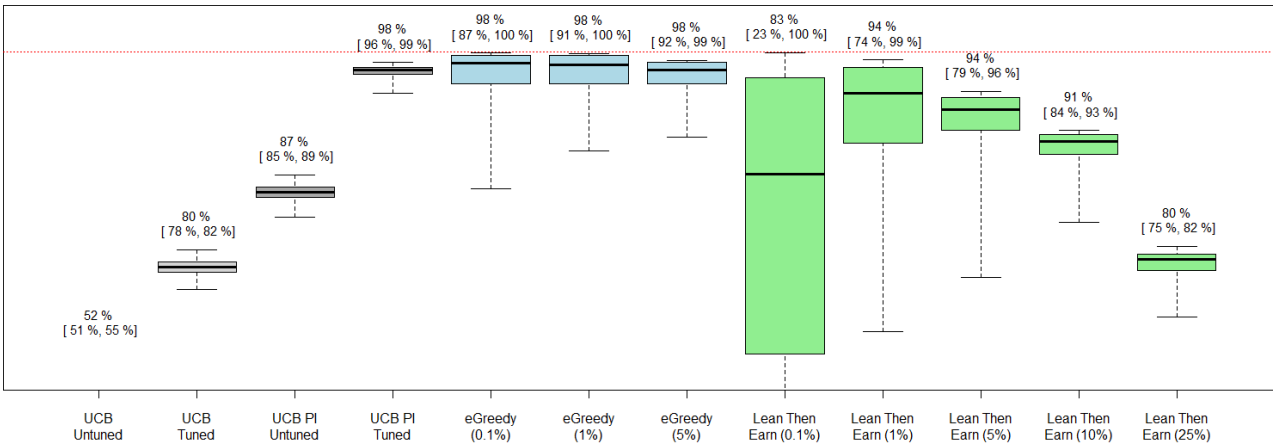
With the emergence of big data, we see an increase in machine learning applications in marketing (Chintagunta et al., 2016). We study a realistic dynamic pricing problem for an online retailer. The goal is a pricing experimentation policy that can apply to many types products (robust) and run in real-time (fast). We propose a novel combination of economic theory with machine learning. To marketing and economics, we bring these scalable reinforcement learning methods to expand the types of dynamic optimization problems. We consider a multi-period dynamic pricing problem, when a firm faces ambiguity. While a two-period solution was available, the method could not realistically scale with time. We propose a fast and scalable algorithm rooted in theory. To the machine learning literature, we introduce distribution-free theory of demand to improve existing algorithms theoretically and empirically. We provide strong evidence for the benefit of partial identification of demand in non-parametric bandit problems.

We note that the advantage of our methodology does depend on the quality of the ex-ante segmentation. In our simulations we considered a value of intra-segment heterogeneity (δ) of 0.1 or 10% of the range of valuations. We repeat our MC simulations for setting with small intra-segment heterogeneity ($\delta = 0.01$) and large intra-segment heterogeneity ($\delta = 0.5$). The results are shown in Figure 4, here we see that as the quality of segmentation (δ) increases, the benefit of partial identification diminishes. With a small value of δ, our method can achieves more than 98% of ex-post optimal profits. However, with a large value of δ our method achieves between only 57% and 91% of ex-post optimal profits. In particular, we tested the maximal value of $\delta = 0.5$, so that the range of every interval was $2\delta = 1$, so the segmentation was meaningless, and all data were pooled at the population level. So it is expected to more closely resemble the typical UCB (without partial identification), as seen in Figure 4, Panel B.

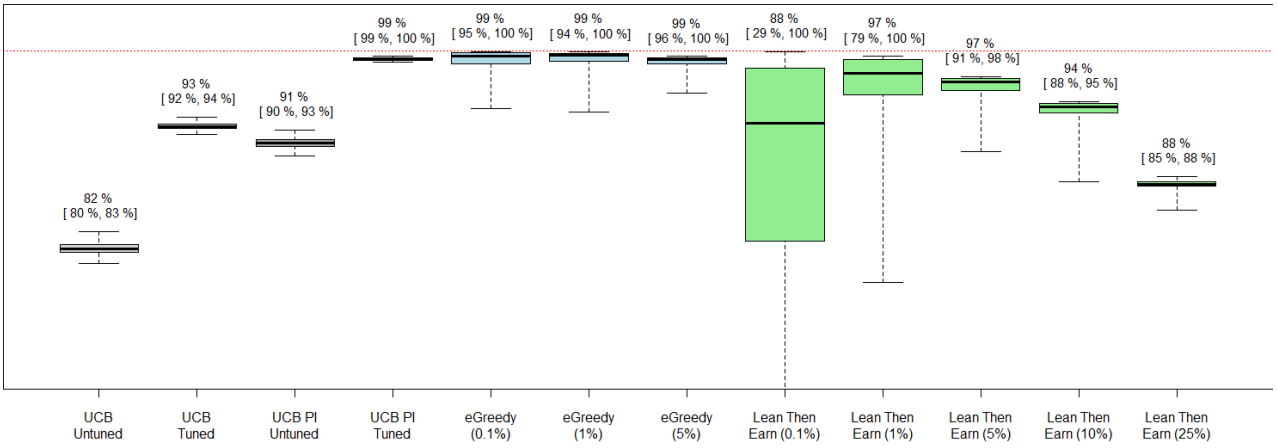
A theoretical limitation of our current work is that we consider a simple demand system. In our model each consumer has a stable valuation. Further research can consider setting where consumer valuation can change over time. This could be in the form of prior prices creating reference prices, or consumer with dynamic preferences.

Panel A: Ex-Post profit achieved in each of our simulation settings. Each chart plots the median (solid line), inter-quintile range (box) and range (whiskers, dotted lines) across 250 Monte Carlo simulations. The numbers represent the mean and the range.

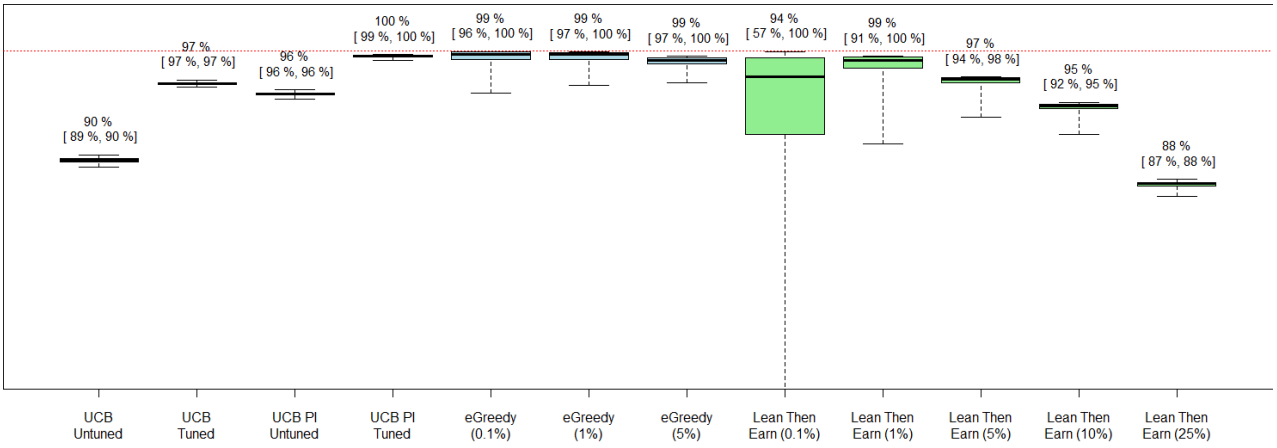
Scenario: Right Skewed



Scenario: Symmetric



Scenario: Left Skewed



Panel B: Range of ex-post optimal profit by algorithm. Chart plots the lowest and highest achieved profit for each of the 12 algorithms, across all 5 simulation scenarios and 250 Monte Carlo experiments.

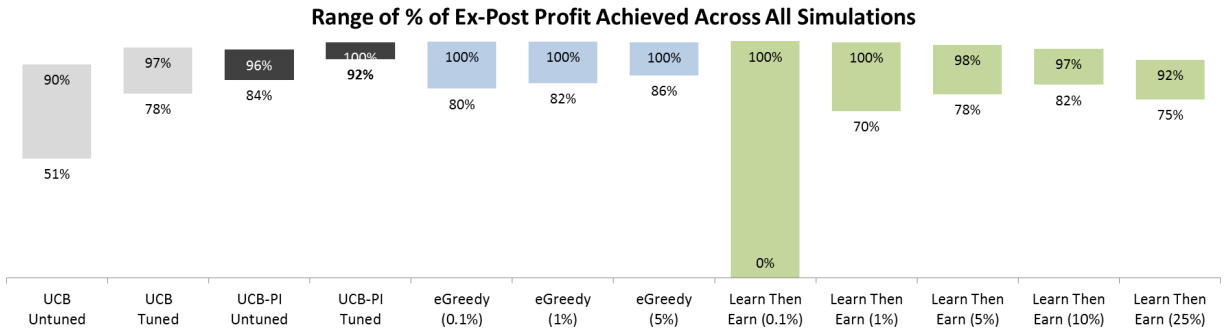
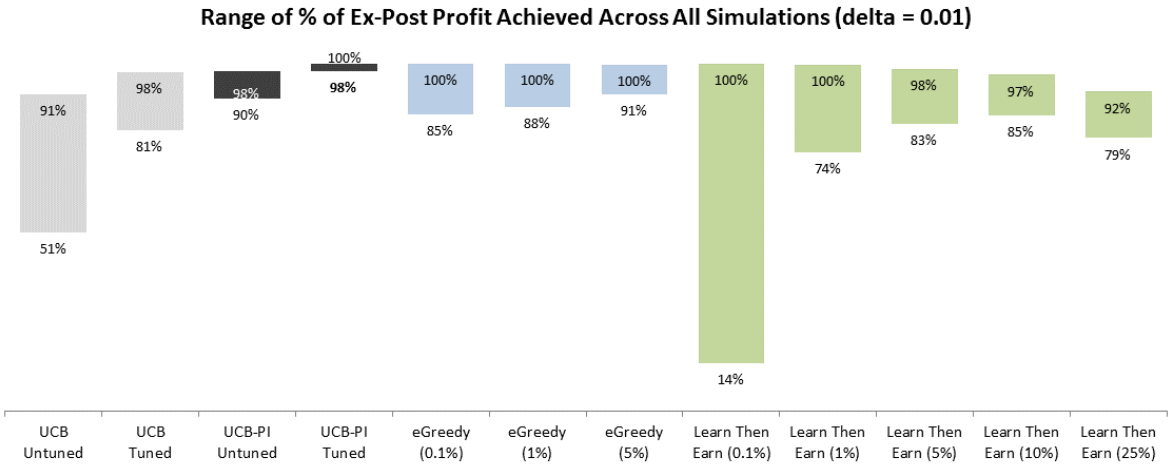


Figure 3: Ex-Post realized profits across 250 Monte Carlo simulations for each our five (5) simulation scenarios.

Panel A: Delta = 0.01 or small intra-segment heterogeneity. Range of ex-post optimal profit by algorithm. Chart plot the lowest and highest achieved profit across all scenarios and Monte Carlo experiments



Panel B: Delta = 0.5 or large intra-segment heterogeneity. Range of ex-post optimal profit by algorithm. Chart plot the lowest and highest achieved profit across all scenarios and Monte Carlo experiments

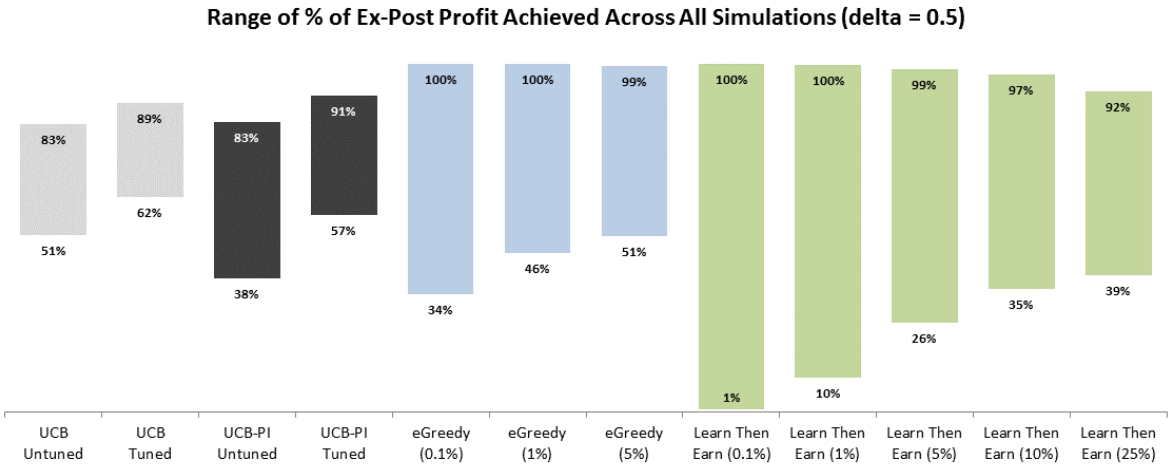


Figure 4: Impact of intra-segment heterogeneity (delta) on the ex-post profits of the algorithm

References

A. Acquisti and H. R. Varian. Conditioning prices on purchase history. *Marketing Science*, 24(3):pp. 367–381, 2005.

P. Aghion, P. Bolton, C. Harris, and B. Jullien. Optimal learning by experimentation. *The Review of Economic Studies*, 58(4):621–654, 1991.

R. Agrawal. Sample Mean Based Index Policies with O(log n) Regret for the Multi-Armed Bandit Problem. *Advances in Applied Probability*, 27(4):1054–1078, 1955.

Y. Akcay, H. P. Natarajan, and S. H. Xu. Joint dynamic pricing of multiple perishable products under consumer choice. *Management Science*, 56(8):pp. 1345–1361, 2010.

E. Anderson, N. Jaimovich, and D. Simester. Price stickiness: Empirical evidence of the menu cost channel. *Review of Economics and Statistics*, 97(4):813–826, 2015.

J. Y. Audibert, R. Munos, and C. Szepesvári. Exploration-Exploitation Trade-Off Using Variance Estimates in Multi-Armed Bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

P. Auer. Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research*, (3):397–422, 2002.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47:235–256, 2002.

Y. Aviv and A. Pazcal. Pricing of short lifce-cycle products through active learning. Unpublished Manuscript, Washington University of St. Louis, October 2002.

B. L. Bayus. The dynamic pricing of next generation consumer durables. *Marketing Science*, 11(3):pp. 251–265, 1992.

D. Bergemann and K. Schlag. Pricing without priors. *Journal of the European Economic Association*, 6 (2-3):560–569, 2008.

D. Bergemann and K. Schlag. Robust monopoly pricing. *Journal of Economic Theory*, 146(6):2527–2543, 2011.

D. Bergemann and J. Valimaki. Market experimentation and pricing. Cowles Foundation Discussion Paper 1122, 4 1996.

J. O. Berger. *Statistical decision theory and Bayesian analysis*. Springer, 1985.

O. Besbes and A. Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420, 2009.

E. Biyalogorsky and E. Gerstner. Contingent pricing to reduce price risks. *Marketing Science*, 23(1):pp. 146–155, 2004.

E. Biyalogorsky and O. Koenigsberg. The design and introduction of product lines when consumer valuations are uncertain. *Production and Operations Management*, 2014.

A. Bonatti. Menu pricing and learning. *American Economic Journal: Microeconomics*, 3(3):124–163, 2011.

D. J. Braden and S. S. Oren. Nonlinear pricing to produce information. *Marketing Science*, 13(3):pp. 310–326, 1994.

M. Brezzi and T. L. Lai. Optimal Learning and Experimentation in Bandit Problems. *Journal of Economic Dynamics and Control*, 27:87–108, 2002.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

P. Chintagunta, D. M. Hanssens, and J. R. Hauser. Editorial—marketing science and big data. *Marketing Science*, 35(3):341–342, 2016.

V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic Linear Optimization Under Bandit Feedback. *Conference on Learning Theory*, 2008.

A. V. den Boer. Surveys in operations research and management science. 2015.

P. S. Desai, O. Koenigsberg, and D. Purohit. Forward buying by retailers. *Journal of Marketing Research*, 47(1):pp. 90–102, 2010.

W. Elmaghraby and P. Keskinocak. Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions. *Management science*, 49(10):1287–1309, 2003.

T. Erdem and M. P. Keane. Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets. *Marketing Science*, 15(1):pp. 1–20, 1996.

S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 586–594. 2010.

J. C. Gittins. *Multi-Armed Bandit Allocation Indices*. John Wiley and Sons, Chichester, UK, 1 edition, 1989.

J. C. Gittins, K. Glazebrook, and R. Weber. *Multi-Armed Bandit Allocation Indices*. John Wiley and Sons, New York, NY, 2 edition, 2011.

P. Hall and B. U. Park. New methods for bias correction at endpoints and boundaries. *Annals of Statistics*, pages 1460–1479, 2002.

B. Handel, K. Misra, and J. Roberts. Robust firm pricing with panel data. *Journal of Econometrics*, 174(2), 2013.

B. R. Handel and K. Misra. Robust new product pricing. *Marketing Science*, 34(6):864–881, 2015.

J. R. Hauser, G. L. Urban, G. Liberali, and M. Braun. Website Morphing. *Marketing Science*, 28(2): 202–223, 2009.

I. Hendel and A. Nevo. Measuring the implications of sales and consumer inventory behavior. *Econometrica*, 74(6):1637–1673, 2006.

G. Hitsch. Optimal dynamic product launch and exit under demand uncertainty. *Marketing Science*, 25(1): pp. 25–30, 2006.

Y. Jiang, J. Shang, C. F. Kemerer, and Y. Liu. Optimizing e-tailer profits and customer savings: Pricing multistage customized online bundles. *Marketing Science*, 30(4):pp. 737–752, 2011.

K. Kalyanam. Pricing decisions under demand uncertainty: A bayesian mixture model approach. *Marketing Science*, 15(3):pp. 207–221, 1996.

G. Kalyanaram and R. S. Winer. Empirical generalizations from reference price research. *Marketing science*, 14(3_supplement):G161–G169, 1995.

R. J. Karunamuni and T. Alberts. On boundary correction in kernel density estimation. *Statistical Methodology*, 2(3):191–212, 2005.

V. Kuleshov and D. Precup. Algorithms for multi-armed bandit problems. *CoRR*, abs/1402.6028, 2014. URL `http://arxiv.org/abs/1402.6028`.

T. L. Lai. Adaptive Treatment Allocation and the Multi-Armed Bandit Problem. *Annals of Statistics*, 15(3): 1091–1114, 1987.

T. L. Lai and H. Robbins. Asymptotically Efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Y. Lei, S. Jasin, and A. Sinha. Near-optimal bisection search for nonparametric dynamic pricing with inventory constraint. 2016.

L. M. Lodish. Applied dynamic pricing and production models with specific application to broadcast spot pricing. *Journal of Marketing Research*, 17(2):pp. 203–211, 1980.

C. Manski. *Social Choice with Partial Knowledge of Treatment Response*. Princton University Press, Princton, 2005.

A. Mas-Colell, M. D. Whinston, J. R. Green, et al. *Microeconomic theory*, volume 1. Oxford university press New York, 1995.

J. Milnor. *Games Against Nature in R.M. Thrall, C.H. Coombs, and R.L. Davis (Eds.) Decision Processes*. Wiley, New York, 1954.

H. Nair. Intertemporal price discrimination with forward-looking consumers: Application to the us market for console video-games. *Quantitative Marketing and Economics*, 5(3):pp. 239–292, 2007.

H. Nair, P. Chintagunta, and J.-P. Dubé. Empirical analysis of indirect network effects in the market for personal digital assistants. *Quantitative Marketing and Economics*, 2(1):23–58, 2004.

S. S. Oren, S. A. Smith, and R. B. Wilson. Nonlinear pricing in markets with interdependent demand. *Marketing Science*, 1(3):pp. 287–313, 1982.

A. Rajan, R. Steinberg, and R. Steinberg. Dynamic pricing and ordering decisions by a monopolist. *Management Science*, 38(2):pp. 240–262, 1992.

R. C. Rao and F. M. Bass. Competition, strategy, and price dynamics: A theoretical and empirical investigation. *Journal of Marketing Research*, 22(3):pp. 283–296, 1985.

M. Rothschild. A two-armed bandit theory of market pricing. *Journal of Economic Theory*, 9(2):185–202, 1974.

P. Rusmevichientong and J. N. Tsitsiklis. Linearly Parameterized Bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

E. M. Schwartz, E. Bradlow, and P. Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Forthcoming in Marketing Science*, 2016.

S. L. Scott. A Modern Bayesian Look at the Multi-Armed Bandit. *Applied Stochastic Models Business and Industry*, 26(6):639–658, 2010.

S. A. Smith. New product pricing in quality sensitive markets. *Marketing Science*, 5(1):pp. 70–87, 1986.

J. Stoye. Axioms for minimax regret choice correspondences. *Journal of Economic Theory*, 146(11): 2226–2251, 2011.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

W. R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3):285–294, 1933.

G. L. Urban, G. Liberali, E. MacDonald, R. Bordley, and J. R. Hauser. Morphing Banner Advertising. *Marketing Science, forthcoming*, 2013.

A. Wald. *Statistical Decision Functions*. Wiley, New York, 1950.

Z. Wang, S. Deng, and Y. Ye. Close the gaps: A learning-while-doing algorithm for single-product revenue management problems. *Operations Research*, 62(2):318–331, 2014.

B. Wernerfelt. A special case of dynamic pricing policy. *Management Science*, 32(12):pp. 1562–1566, 1986.

White House. Big data and differential pricing. 2015. URL `https://www.whitehouse.gov/sites/default/files/docs/Big_Data_Report_Nonembargo_v2.pdf`.

P. Whittle. Multi-armed Bandits and the Gittins Index. *Journal of Royal Statistical Society, Series B*, 42(2): 143–149, 1980.

R. S. Winer. A reference price model of brand choice for frequently purchased products. *Journal of consumer research*, pages 250–256, 1986.